Colorful Language

A Presentation About “Distributional Semantics in Technicolor”

Jetze Baumfalk

ILLC
University of Amsterdam

Computational Semantics, 2012

Outline

Goal: Connecting Language and Perception

- Traditional semantic space models represent meaning on the basis of text corpora.
- Human semantic knowledge relies on non-verbal experience and different representations (e.g. vision).
- Can we use this to build better semantic models?

Approach to finding an answer

- Compare models using textual, visual and both types of information.
- Evaluate models on general semantic relatedness task and visual-sensitive tasks.
- Compare different kinds of visual models.

Textual Models

Four textual models where used:

1. **Window2**: nearest sentence-internal co-occurrence with two word window.
2. **Window20**: nearest sentence-internal co-occurrence with twenty word window.
3. **Document**: “topic-based” approach, words are represented as distributions over documents.
4. **Distributional Memory**: exploits lexico-syntactic and dependency relations. It is a grammar based model.

All models used the ukWac and the Wackypedia corpora (3B tokens combined). In addition, the Distributional Model also used the BNC corpus.

Visual Models - Image Data

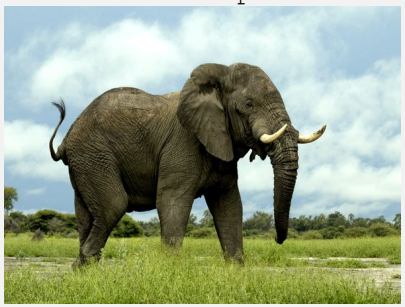
Each image of the image date was associated with one or more tags. The set of tags is called the label of the image. For example,

could have the label *{elephant, savanna}*.

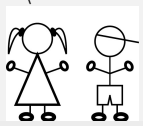
Visual Models - ESP Game

The images where labeled using the ESP Game. It works as follows. Two people had to agree on a tag for an image:

`Label this picture`



“Elephant” “Elephant”

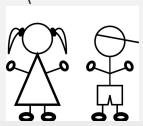


Visual Models - ESP Game (2)

The second pair of people had to do the same thing, but couldn’t use elephant.

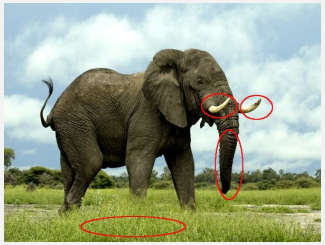
`Label this picture (but no elephant)`

“Savanna” “Savanna”



Visual Models - Visual Features

For each tag, a vector was built with visual features. These visual features were extracted using bag-of-visual-words (*BoVW*). First, relevant areas are identified.



Then, a low-level feature vector (called a *descriptor*) is built to represent each area.

Visual Models - Visual Features (2)

This descriptor, who lives in a "descriptor space", are grouped into k clusters. Each cluster can be seen as a visual word. The tags of the image are represented by the vector with these k clusters as dimensions.

Tag	"Grass"	"Tusk"	"Trunk"
Elephant	1	2	1
Savanna	1	2	1

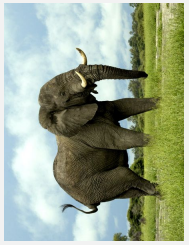
And after many more images:

Tag	"Grass"	"Tusk"	"Trunk"
Elephant	50	600	700
Savanna	1000	30	15

Visual Models - Descriptor Features

Two descriptor features are extracted.

- Scale-Invariant Feature Transform (SIFT) feature vectors.
  - Good at characterizing parts of objects
  - Works independent to image scale and rotation.


=


  - k visual words, $k = 500, 1000, \ldots, 2500$
- LAB feature vectors.
  - A way of depicting colors (like RGB).
  - k visual words, $k = 128, 256, \ldots, 1024$

Multimodal Models

Multimodal models are simply the vectors of the textual and visual models concatenated.

The "multimodal" vector F is calculated by $\alpha \times F_t \oplus (1 - \alpha) \times F_v$. Here F_t is the vector of the textual model, F_v is the vector of the visual model and $\oplus$ is vector concatenation.

Hybrid models

Hybrid models are similar to the two other models:

- Similar to textual models since they are based on co-occurrence.
- Similar to visual models since they consider co-occurrence in the image labels.

There are two kinds of hybrid models:

- ESP-Win: like window based models.
- ESP-Doc: like document based models.

Evaluation as General Semantic Models

For the evaluation, two datasets where used.

1. The WordSim353 dataset, contains similarity between words. Range: [0,10].
   - "*dollar*/*buck*" = 9.22
   - "*professor*/*cucumber*" = 0.31
2. The (new) MEN dataset, contains relatedness between words. Range: [0,1].
   - "*cold*/*frost*" = 0.9
   - "*eat*/*hair*" = 0.1

Experiment 1: The Color Of Concrete Objects

Hypothesis: Relation between the words for concrete objects and their color is reflected better by visual models.

Testing:

- Label list of concrete nouns (*"grass"*, *"crow"*) with their typical color (*"green"*, *"black"*).
- Measure the cosine ("relatedness") between the noun and the color vector produced by the models.

Experiment 2: Literal Versus Nonliteral Color Uses

Hypothesis: Literal color (*"green grass"*) usages in a good model will have higher similarity between the noun and the color term.

Testing:

- Generate list of color noun pairs (*"green grass"*, *"red district"*).
- Label the pairs as literal or nonliteral.
- Measure the average cosine between noun and color across literal and nonliteral pairs.

Results From The Experiments

Model	*WS*	*MEN*	*E1*	*E2*
DM	.44	.42	3 (09)	.14
Document	.63	.62	3 (07)	.06
Window2	**.70**	.66	5 (13)	.49***
Window20	**.70**	.62	3 (11)	.53***
LAB_{128}	.21	.41	**1** (27)	.25*
LAB_{256}	.21	.41	2 (24)	.24*
LAB_{1024}	.19	.41	2 (24)	.28**
$SIFT_{2.5K}$	.33	.44	3 (15)	.57***
W2-LAB_{128}	.40	.59	**1** (27)	.40***
W2-LAB_{256}	.41	.60	2 (23)	.40***
W2-LAB_{1024}	.39	.61	2 (24)	.44***
W20-LAB_{128}	.40	.60	**1** (27)	.36***
W20-LAB_{256}	.41	.60	2 (23)	.36***
W20-LAB_{1024}	.39	.62	2 (24)	.40***
W2-$SIFT_{2.5K}$	.64	**.69**	2.5 (19)	.68***
W20-$SIFT_{2.5K}$	.64	.68	2 (17)	**.73****
ESP-Doc	.52	.66	**1** (37)	.29*
ESP-Win	.55	.68	4 (15)	.16

Interpretation Of The Results

- In a similarity task, a text only model is the best option.
- In a relatedness task, using some visual information yields a better model.
- Shallow visual models can deal with easy visual oriented tasks.
- Multimodal models are better for sophisticated tasks.

Summary/Discussion

Take-home messages

- Models using visual information can be better than text based models for the visual semantic aspect of words.
- Computer vision is getting mature enough to contribute significantly to perceptually grounded computational models of language.

Discussion Questions

- What other visual information could be used to improve the model?
- Are there other modalities (sound, touch) that can contribute to semantic models?